Executive Steering Committee
For A.C.E. Policy II
(ESCAP II)
Report No. 21

September 24, 2001

ESCAP II: Analysis of Census Imputations

Fay F. Nash

Decennial Management Division

USCENSUSBUREAU

Helping You Make Informed Decisions

CONTENTS

EX	KECUTIVE SUMMARY	ii
1.	BACKGROUND	. 1
2.	COUNT IMPUTATIONS 2.1 Background 2.2 Results 2.3 Conclusion	. 1
3.	WHOLE PERSON CHARACTERISTICS IMPUTATION 3.1 Background 3.2 Results 3.3 Conclusion	. 5
Αp	opendix A-1: Count Imputation Rates in the Decennial Census	. 8

EXECUTIVE SUMMARY

Census 2000 experienced a higher rate of whole person imputations than in the 1990 census. Whole person imputations are excluded from A.C.E. matching activities, but are reflected in the census coverage error as measured by the A.C.E. dual system estimates. This report provides information as to the sources of the whole person imputations.

What conclusions should be drawn as a result of this analysis with respect to the adjustment decision?

The kind, level and pattern of whole person imputation in Census 2000 raise no additional issues relative to the accuracy of the A.C.E. adjustment.

What were the sources of the whole person imputations?

Approximately 5.77 million persons had all their characteristics (hundred percent data items) imputed in Census 2000 (compared to 1.97 million persons in the 1990 census). Approximately 1.2 million of these persons were added to the census count through a count imputation process. The remaining 4.6 million persons were counted directly through the census enumeration process, but had all their person characteristics data imputed because information about them was substantially missing from the census records.

- C The count imputation process adds persons to the census. For these imputed persons, their characteristics are necessarily imputed also. In Census 2000, 1,172,144 persons were imputed through count imputation, or 0.42 percent of the total population. While this rate was in line with earlier censuses, it was higher than the rate experienced in1990, as depicted by the attached chart.
- C The whole person characteristics imputation process supplies all the characteristics data for persons already counted directly through the census enumeration process. In Census 2000, 4,602,122 persons were imputed through the whole person characteristics imputation process, or 1.64 percent of the total population. This was also a higher rate than the 1990 census rate of 0.77 percent.

There were five basic categories of census cases that required an imputation procedure to fill in missing data as to the person counts and/or all (hundred percent) person characteristics items - three are categories of the count imputation process; two, of the whole person characteristics imputation process. The results for all five categories are summarized in the following table.

Table 1. Whole Person Imputations by Category

	Housing Units Requiring Imputation	Imputed Persons
Total Count Imputations	620,650*	1,172,144
1. Status Imputation—No information about the housing unit	235,071*	415,892
Occupancy Imputation – Existence of housing unit confirmed; no information as to occupancy status	191,826	260,652
 Household Size Imputation — Occupied status confirmed; no information as to household count 	193,753	495,600
Total Whole Person Characteristics Imputations	2,261,664	4,602,122
4. Whole Household Imputation– Population count known; all characteristics imputed for the entire household	1,006,111	2,269,010
5. Within Household Imputation– Population count known; all characteristics imputed for some, but not all, persons in the household	1,255,553	2,333,112
Total Whole Person Imputations	2,882,314	5,774,266

^{*} Excludes cases with an imputed "Delete" status and thus removed from the census.

What Census 2000 design features or processes contributed to the count imputation rate?

Reasons for the increase in count imputations when compared to 1990 are explainable. Some of the increase was expected due to changes in the census design. The reasons vary by category.

C Status Imputation -- Seventy-five (75) percent of the cases were housing units added to the Census Master Address File (MAF) after the census mailing/delivery list was produced. They reflect valid housing units added either by enumerators in the field or by respondents themselves. Such adds are introduced into the process via census questionnaires that contain no MAF ID numbers. Once MAF ID numbers are assigned to these added housing units, the processing system merges the addresses/IDs with their corresponding questionnaire data via temporary processing numbers assigned earlier. Two situations occurred that prevented the Census Bureau from linking some number of housing units back to their corresponding data and thus causing them to undergo the status imputation process. In one situation, the temporary processing number occasionally became corrupted, causing the link between the MAF ID and the questionnaire data to be broken. In the other situation, some of the MAF IDs were assigned late in the processing cycle and were consequently not added to the census until after the headquarters process of merging data to MAF IDs had occurred.

The remainder of the status imputation cases were primarily the result of census records that were data captured, but which contained no data; i.e., blank census records.

C Occupancy Imputation -- Ninety-three (93) percent of the cases were enumerator-completed questionnaires with inconsistent data recorded on them that precluded a definitive classification of either occupied or vacant. In Census 2000 no clerical edit process was implemented to resolve such cases as was done in the 1990 census. Instead, such inconsistencies were resolved via the occupancy imputation process, leading to a more standardized approach. That is, cases requiring occupancy imputation were expected because of this change in the Census 2000 design.

The remaining cases were cases confirmed to be valid housing units, but for which no questionnaire data could be linked or the census record captured was blank.

C Household Size Imputation -- Eighty-three (83) percent of the cases were enumerator-completed cases with clear information to classify the case as occupied, but with missing or inconsistent data as to household size. As with occupancy imputation, these cases were expected because of the census design change that instituted a standardized imputation process, rather than a clerical edit procedure.

The remaining cases were primarily instances where the only forms data captured for the MAF IDs were specialized, single-person data collection forms. As single-person forms do not reflect overall household size, household size is necessarily imputed.

What Census 2000 design features or processes contributed to the whole person characteristics imputation rate?

Reasons for the increase in whole person characteristics imputations when compared to 1990 are explainable. Some of the increase was expected due to changes in the census design. The reasons vary by category.

- Whole Household Imputation -- Eighty-nine (89) percent were enumerator-completed questionnaires, 8 percent were mail return questionnaires, and 2 percent had no questionnaire data that could be linked to the ID or the census record captured was blank. The rate of whole household imputations for person characteristics in Census 2000 is comparable to the 1990 census rate so should not be a matter of concern.
- Within Household Imputation -- Sixty-seven (67) percent were from mail-returns. These mostly reflect the large households (size seven or more) that were not accommodated by the 6-person mail-return questionnaire and for which the telephone followup operation was unable to supply the missing data. More persons were expected to fall into this imputation category than in 1990 because the Census 2000 mail-return questionnaire accommodated fewer persons than the 1990 form (that is, six persons rather than seven), in order to implement a more user-friendly, easily data-captured questionnaire.

The remaining cases were from enumerator-completed questionnaires, where 1) the enumerator was unable to collect information about all household members or 2) the continuation form was unable to be linked to its parent form for some large households.

1. BACKGROUND

Approximately 5.77 million persons were whole person imputations in Census 2000. That is, they had all their characteristics (sex, age/date of birth, relationship, Hispanic origin and race) imputed. Approximately 1.2 million of these persons were added to the census count through a count imputation process. Persons added through the count imputation process necessarily had their person characteristics data imputed. The remaining 4.6 million persons were counted directly through the census enumeration process, but had all their person characteristics data imputed because this information was substantially missing from the census records.

This study provides information as to the sources of the whole person imputations. The data presented are the result of research conducted by an interdivsional team tasked to investigate the reasons why a higher rate was experienced in Census 2000 when compared to the 1990 census. It is divided into two main sections - one that describes the imputations resulting from the count imputation process and the other that describes the imputations resulting from the whole person characteristics imputation process.

2. COUNT IMPUTATION

2.1 Background

A total of 1,172,144 persons, or 0.42 percent of the total population, was added to the apportionment count in Census 2000 through count imputation. While this rate was in line with earlier censuses, it was higher than the rate of count imputation in the 1990 Census (see attached chart).

The Census Bureau used count imputation in Census 2000 as it has in several prior censuses to address the problem of missing, incomplete, and contradictory data. The Census Bureau used count imputation for three categories of cases in Census 2000:

- <u>Household Size Imputation</u> The Census Bureau imputed a population count for a housing unit when Bureau records indicated that the housing unit was occupied, but had insufficient information as to the number of individuals residing in the unit.
- Occupancy Imputation When Census Bureau records indicated that a housing unit existed but did not provide sufficient information to definitively classify it as either occupied or vacant, the Bureau imputed occupancy status (occupied or vacant). For a unit imputed as occupied, household size was also imputed.
- <u>Status Imputation</u> When the Census Bureau's records had insufficient information about whether an address represented a valid, non-duplicated housing unit, the Bureau imputed the status of the unit (occupied, vacant, or delete). For a unit imputed as occupied, household size was also imputed.

2.2 Results

As is shown in Table 2, the number of housing units subject to each of these three categories of count imputation was roughly equal.

Table 2. Census 2000 Count Imputation by Category

	Housing Units Requiring Imputation	Imputed Persons
Total	620,650*	1,172,144
Status Imputation	235,071* (38%)	415,892 (36%)
Occupancy Imputation	191,826 (31%)	260,652 (22%)
Household Size Imputation	193,753 (31%)	495,600 (42%)

^{*} Excludes cases with an imputed "Delete" status and thus removed from the census.

The explanations as to why more housing units were handled by the imputation process in Census 2000 than in 1990 vary by category of count imputation.

2.2.1 Explanation for the number of status imputations

Status imputation in Census 2000 contributed to 235,071 housing units imputed as occupied or vacant, resulting in the imputation of 415,892 persons. The vast majority of the imputed housing units (97%) were No Return cases, with the remaining 3% being Enumerator Return cases. The Enumerator Return cases required status imputation because the questionnaires contained inconsistent information as to whether the unit should be classified as occupied, vacant or delete.

The No Return cases were those cases which were included on the Decennial Master Address File (DMAF) at the end of the census, but for which no data record was associated. Research revealed that 176,832 units (75% of the entire status imputation category) were census adds, meaning that they were housing units that were not on the DMAF by the time questionnaires were mailed out or delivered, but rather were added either by enumerators during field operations or by respondents themselves.

There are two reasons why we have no data for these added cases - the Non Identification (Non-ID) process and the constraints on the data processing schedule. The Non-ID process is a method by which new addresses are added to the census address file. A unique identifier, the MAF Identification (ID), is preprinted on questionnaires mailed or delivered to addresses for mailback, and preprinted on questionnaires given to enumerators for addresses they are assigned to visit during the field followup operations. However, enumerators working in the field may find previously missed addresses that need to be added to the census - addresses that do not have MAF IDs pre-assigned. Additionally, respondents may have completed Be Counted forms or responded through telephone questionnaire assistance (TQA), both of which would result in

respondent data without a MAF ID. These added cases are assigned a temporary processing ID. The Non-ID process then matches these addresses to the MAF and assigns a MAF ID. The temporary processing ID links the newly assigned MAF ID to the appropriate data capture record. For some cases, the temporary processing ID was corrupted on the transmittal file of adds that entered the Non-ID process, thus preventing us from linking back to the corresponding data capture record.

Constraints on the data processing schedule were also a source of missing data for census adds. Some MAF IDs were entered on the DMAF very late in the census process - after the headquarters processing activities had begun. This is particularly true for the census operations conducted later in the census schedule, such as for the Coverage Improvement Followup (CIFU) operation. These late additions were not included in the merge process, whereby data captured records were merged with corresponding MAF IDs, that occurred at the beginning of the headquarters processing. Consequently, no data records were associated with these late adds/MAF IDs. We only discovered this problem after the census counts had been released.

In summary, 75% of the housing units imputed in the status imputation category were added during the census enumeration process. They reflect valid housing units that were added either by enumerators in the field or by respondents themselves. These cases required imputation because we could not associate the corresponding data records to their final census ID numbers. However, these cases were appropriately included in the census.

The remaining 50,674 No Return cases (comprising 22% of the total status imputation category) consisted of census records that were data captured, but which contained no data, i.e., blank census records.

2.2.2 Explanation for the number of occupancy imputations

Occupancy imputation in Census 2000 contributed to 191,826 housing units imputed as occupied or vacant, resulting in the imputation of 260,652 persons. Of the housing units that required occupancy imputation, 179,149 (93%) were Enumerator Returns, 12,175 (6%) were in the No Return category, with the remaining 502 (0%) comprising the Mail Returns.

The Enumerator Return cases required imputation because of inconsistent data recorded on the questionnaires, which precluded a definitive classification of either occupied or vacant. For example, the interviewer summary items may indicate the unit is vacant, but person data is recorded on the questionnaire. In Census 2000, no clerical edit process was implemented to resolve such inconsistencies prior to data capture as was done in the 1990 census. Instead, interviewer inconsistencies were handled by assigning an occupancy status via the automated imputation process, leading to a more standardized process.

The No Return cases were determined to be census adds verified in the Field Verification operation to exist as separate housing units, but for which no data capture record could be associated or for which the only record data captured was blank. The Non-ID process and

constraints on the data processing schedule contributed to this group of cases, as well as to those cases described under the status imputation category (see above for more detail).

2.2.3 Explanation for the number of household size imputations

Household size imputation contributed to 193,753 occupied housing units in Census 2000 with population counts imputed, resulting in the imputation of 495,600 persons for this category. Of the housing units that were imputed, 159,761 (83%) were from Enumerator Returns, 29,402 (15%) were from Specialized Returns (such as Individual Census Returns, Individual Census Questionnaires, Military Census Returns, and Shipboard Census Returns), and 4,590 (2%) were in the No Return category.

The Enumerator Return cases required imputation because although the census record clearly indicated the unit was occupied, there was insufficient information about the household size. As with the Enumerator Returns requiring occupancy imputation, the higher rate of cases imputed under this category than in the 1990 Census can be substantially explained by the fact that for Census 2000 we did not perform a clerical coverage edit prior to data capture as was done in the 1990 Census.¹ Inconsistent or missing data caused these cases to be included in the count imputation process.

The Specialized Returns are single-person data collection forms. When these are the only forms data captured for a MAF ID, the ID does not necessarily represent a single-person household. Consequently, household size is imputed for such housing units.

The No Return cases are those for which the occupied status has been verified through a field operation, but for which no information is available from a census data record, i.e., a blank census record.

2.3 Conclusion

Most of the count imputations performed in Census 2000 are attributable to housing units that have been determined to exist, but whose data were not included in the totals through a variety of reasons. These cases have been appropriately included in the census. If they had not been included in the count imputation process, these cases would represent individuals or housing units that should have been included in the Census, but who were left out because of incomplete or inconsistent data or the inability to locate appropriate data records due to processing system issues.

¹The number of this variety of count imputations would have been higher but for the fact that during the mid stages of the Nonresponse Followup operation, the Census Bureau discovered that a fairly large percentage of questionnaires were being processed with no population count. In order to reduce the number of potential person imputations, we implemented a process to identify these cases and send them back to the field to retrieve the missing information.

3. WHOLE PERSON CHARACTERISTICS IMPUTATION

3.1 Background

The Census Bureau also used an imputation process to complete all the person characteristics data for an individual when such data is substantially missing from the census record. These persons were counted directly by the census enumeration process, but are missing data about them. Persons do not undergo a whole person characteristics imputation process if their census records contain two or more of the 100% population data items or name. (Such persons are called data-defined persons.) The Census Bureau imputed whole person characteristics for two categories of cases in Census 2000. These categories employ different imputation methodologies.

- Whole Household Imputation -- These households contain no data-defined persons and thus require all characteristics data to be imputed for each of the household members. The imputation process replicates all of the 100% person data items (sex, age/date of birth, relationship, Hispanic origin, and race) by substituting data from a hot deck nearest neighbor donor pool record of the same household size.
- Within Household Imputation -- These households contain at least one data-defined person, but other persons with substantially missing data. The imputation process allocates missing values for individual person characteristics data items on the basis of other reported information for the person or household, or from other persons or households with similar characteristics.

3.2 Results

As shown in Table 3, the number of housing units subject to each of these categories is roughly equal.

Table 3. Census 2000 Whole Person Characteristics Imputation by Category

	Housing Units Requiring Imputation	Imputed Persons	
Total	2,261,664	4,602,122	
Whole Household Imputation	1,006,111 (44%)	2,269,010 (49%)	
Within Household Imputation	1,255,553 (56%)	2,333,112 (51%)	

Explanations as to why housing units were required to undergo a whole person characteristics imputation process are discussed by category.

3.2.1 Explanation for the number of whole household imputations

There were 1,006,111 whole household imputations, resulting in 2,269,010 persons with all person characteristics imputed. This is comparable to the 1990 census experience. Of the households that were imputed, 899,295 (89%) were from Enumerator Returns, 84,927 (8%) were from Mail Returns, 203 (0%) were from Specialized Returns (such as Individual Census Returns) and 21,686 (2%) were in the No Return category.

The Enumerator Return cases comprise the vast majority of cases requiring person characteristics data to be imputed for all persons in the household. As with all census operations, the field data collection operations must end on schedule in order for the Census Bureau to meet its legislatively mandated data delivery dates. Toward the end of the Nonresponse Followup operation, an intense and concentrated effort was instituted to quickly and efficiently obtain census data for the last 5 percent of the difficult, unresolved cases. Enumerators were required to obtain at least the housing unit status and the population count for these cases. When these were the only items an enumerator could obtain, the household became a candidate for the whole household imputation process.

The Mail Return cases reflect those households that chose to provide the Census Bureau with only the most minimal data - the population count.

The No Return cases were blank Enumerator Returns. For these cases, the population count was obtained from the Field Division's office control system (the OCS2000).

3.2.2 Explanation for the number of within household imputations

There were 1,255,553 within household imputations, resulting in 2,333,112 persons with all person characteristics imputed. Of the households with imputed person data, 845,187 (67%) were from Mail Returns, 400,539 (32%) were from Enumerator Returns, and 9,827 (1%) were from Specialized Returns (such as Be Counted questionnaires).

The Mail Return households were primarily cases that were directed to the Coverage Edit Followup (CEFU) operation. This operation was designed as a telephone followup for certain mail return households to improve data quality and within household coverage. All households of size 7 or greater were directed to the operation to collect person characteristics data for those persons not accommodated by the 6-person questionnaire. Other cases were directed for followup to resolve population count discrepancies between the reported household population count and the actual number of person data records on the census mail back form. If the telephone followup was not able to obtain the additional person data for an individual in the household, the person characteristics data were imputed. By design, more households were candidates for this telephone followup than in the 1990 census because the 2000 mail return

questionnaire only accommodated six persons, where as the 1990 form accommodated seven. In addition, it is possible that response rate for CEFU was less than that achieved during the 1990 census followup due to the greater reluctance of households to provide telephone numbers, the increased use of answering machines and caller ID technology to screen calls or other reasons. This cannot be verified because comparable performance measures are not available from the 1990 followup operation.

The Enumerator Return cases included those for which enumerators were unable to obtain person characteristics data for all the household members. In addition, some continuation forms for large households could not be linked back to their parent forms during headquarters processing. This occurred when the continuation forms were not inserted inside the parent forms at the time the forms were checked into the Data Capture Centers. If the forms became separated and the continuation form lacked sufficient address information to allow it to be subsequently linked back to the parent form, the person data on the continuation form became lost. The person characteristics data would consequently be imputed. This occurred primarily in list-enumerate areas.

The Specialized Return cases were primarily Be Counted questionnaires where respondents completed data for some persons in the household, but not all. Also included in this category are large households (size 6 or greater) responding via the 5-person Be Counted questionnaire, but for which the CEFU operation was unable to obtain the additional person characteristics information. (See discussion for mail returns for more detail.)

3.3 Conclusion

The increase in whole person characteristics imputations in Census 2000 as compared to the 1990 census is primarily attributable to the within household imputation category. The increase in this category is explained by changes in census design features, as well as the increasing difficulty in collecting census data for all persons.

Count Imputation Rates in the Decennial Census

